censusbrowseR: A User-Friendly U.S. Census Browser for R

Kiegan Rice, *Iowa State University*

Introduction

Census data provide an important snapshot of information about a country at different times throughout its history, and their value is difficult to overstate. While history books present a narrative of the events that occured, students themselves often don't get to interact with the raw data in that learning environment. Clean and accessible census data allow the exploration of different demographic groups over time, or investigations of a particular period of time and what the demographic and economic landscape looked like in the past. Even today, data about the world around us opens a pathway for learning more about places we haven't been and groups of people with whom we may not usually engage.

From an early point in the United States' history, there were many "eminent men of science" who recognized the value of the census data and worked to aggregate and present those data. Francis A. Walker's "Statistical Atlas of the United States", based on the 1870 census, was an impressive effort in aggregating population data to present them in a visually appealing way [3]. Although the Census Bureau's "Statistical Atlases" eventually stopped being created, they were an important start to the effort of visually presenting census data to a wider public (see Figure 1).

Figure 1: *A visualization of the African American population in the United States from the "Statistical Atlas of the United States".*

Today, as methods of data analysis and visualization continue to be developed and improved, access to census data allows us to look back on that history and explore, synthesize, and visualize the information. When aggregated and presented in a clear manner, viewers can learn more about patterns in many different parts of the population. A 2007 paper about the Statistical Atlas discusses just how much information is present in those original charts and demonstrates several other ways of presenting the information that can be found therein [2].

Ever-improving visualization and data-wrangling methods in R [4] give those interested in statistical graphics a wealth of opportunities to explore and learn from data; in particular, incorporating user interactivity using Shiny has revolutionized the way statisticians share and communicate information [1]. However, it is difficult to make use of these tools with census data if those data are not available and easy to explore in one location.

An inherent problem in census data is that a country's census changes over time; the variables collected, how they are collected, and even the locations on which they are collected are updated as the country is formed, and subsequently grows and changes. The United States census data are no exception to this rule. In little under two and a half centuries, the census has taken many different forms. Data on occupations have transformed as the employment landscape has changed; new states have been formed, the most recent being within the last 100 years; definitions of various demographic groups and the terminology used to describe them have been updated as the demographic makeup of the country has changed. Each decennial census brings a different set of variables to the table. Sometimes these variables are new things the Census Bureau is interested in learning about, while sometimes they remove variables that are no longer relevant or whose information is captured elsewhere.

Unfortunately, because the founders of the U.S. Census were unable to see 200 years into the future, those interested in working with census data are left with quite the inescapable mess. If you want to focus in on a particular demographic group and their journey as part of the population of the United States, you may have ten or more different variable names to describe that one group over the course of the census from 1790 to 1960 - and that is just for a single group! Of course, we cannot just simply change variable names to match our own research needs. It is important to keep the data in their true form and be respectful of the way in which the population was defined at different times throughout history, even if our instinct may be to 'clean' the data by changing variable names.

This, of course, leaves the user with a wide variety of variables that are far from consistent across years. In order to track one demographic group across years - let alone many groups - a clean user interface that helps users see exactly what information is available to them is a necessity. To streamline the process and assist researchers in finding out to what information they have access and what information they lack, a U.S. Census Browser for R is presented as an R package, with the user interface being a Shiny application, and the downloadable files being 'tidy' comma-separated values (csv) files that those with a minimal amount of R experience can still process.

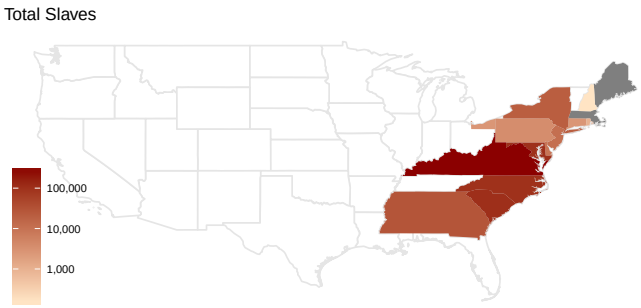

Figure 2: *Total number of slaves per state in 1790, plotted on a continuous log scale. State boundaries for July 4, 1790 were gathered from 'USAboundaries' package.*

Data Access

The `censusbrowseR` package

After ordering by how many years each variable appears in, the available variable names show a serious lack of continuity. The four ever-present variables (`Year`, `State`, `TOTAL.POPULATION`, and `Type`) each appear ten times, once in each year selected. After that, the next most common variable only appears in four of the ten years we have chosen. This highlights some of the hurdles that we will encounter when tracking any demographic group over time.

Since the focus will be on African Americans throughout U.S. history, the first search term needed is `SLAVE` in the early years of the U.S. census. The vast majority of African Americans were slaves when the United States was founded, so variables that count the number of slaves tell the story about the African American population in early U.S. history. Note that the 'SLAVE' categorization is the only term the U.S. census had related to African Americans in the early years of the census, so it is also the only source of information available on that group for that time period.

Blah blah blah Figure 2 is important.

Conclusion

The decennial census is a rich source of historical information. We have built a tool that can help extract and visualize that information to, for example, understand how demographic attributes of the U.S. population have changed over time.

Although there is no county-level information available and the data for the census browser go only through 1960, users are able to efficiently explore variables of interest at different time points in U.S. history and tell a visual story about a growing nation and a changing population. This connectivity across years and ability to assess what information was available when, all in one browser, have the potential to save time and effort of researchers interested in delving into United States history.

The structure of our browsers also helps determine not only the variables that are available in a given year, but also where information is missing for some states on a particular variable. Finding these differences is easier when all relevant variables are combined in one data frame, which can be summarized and searched all at one time.

References

[1] W. Chang, J. Cheng, J. Allaire, Y. Xie, and J. McPherson. shiny: Web application framework for r., 2017.

[2] H. Hofmann. Interview with a centennial chart. *CHANCE*, 20:2:26–35, 2007.

[3] U. S. C. Office and F. A. Walker. *Statistical atlas of the United States based on the results of the ninth census 1870 with contributions from many eminent men of science and several departments of the government.* 1874.

[4] R Core Team. R: A language and environment for statistical computing. 2015.

Further Reading